Lecture 10: Maximum likelihood IV. (nonlinear least square fits)

$\chi^2$ fitting procedure!
An example might be something like fitting a known functional form to data

\[ f(x) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2} \frac{(x - b_4)^2}{b_5^2}\right) = 2 \cdot p(x) - 0.4 = y(x|b) \]

measured value of 2p-0.4 as a function of x

For example, this rise might be an instrumental or noise effect, while this bump might be what you are really interested in.
from Lecture 9:  Maximum Likelihood discussion

Fitting is usually presented in frequentist, MLE language. But one can equally well think of it as Bayesian:

\[ P(b|\{y_i\}) \propto P(\{y_i\}|b)P(b) \]
\[ \propto \prod_i \exp \left[ -\frac{1}{2} \left( \frac{y_i - y(x_i|b)}{\sigma_i} \right)^2 \right] P(b) \]
\[ \propto \exp \left[ -\frac{1}{2} \sum_i \left( \frac{y_i - y(x_i|b)}{\sigma_i} \right)^2 \right] P(b) \]
\[ \propto \exp \left[ -\frac{1}{2} \chi^2(b) \right] P(b) \]

frequentist: \( P(b) \sim \delta(b-b_0) \quad b_0? \)

Bayesian: \( P(b) \sim \text{const} \quad \) simplest, leads to same \( b_0 \) determination

repeating the experiment with \( y_i \) and \( \sigma_i \) we also test \( f(x) \) as a hypothesis

Now the idea is: Find (somehow!) the parameter value \( b_0 \) that minimizes \( \chi^2 \).

For linear models, you can solve linear “normal equations” or, better, use Singular Value Decomposition. See NR3 section 15.4

In the general nonlinear case, you have a general minimization problem, for which there are various algorithms, none perfect.

Those parameters are the MLE. (So it is Bayes with uniform prior.)
Nonlinear fits are often easy in MATLAB (or other high-level languages) if you can make a reasonable starting guess for the parameters:

\[
y(x|b) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2} \left(\frac{x - b_4}{b_5^2}\right)^2\right)
\]

\[
\chi^2 = \sum_i \left(\frac{y_i - y(x_i|b)}{\sigma_i}\right)^2
\]

```
ymodel = @(x,b) b(1)*exp(-b(2)*x)+b(3)*exp(-(1/2)*((x-b(4))/b(5)).^2)
chisqfun = @(b) sum(((ymodel(x,b)-y)./sigma).^2)
```

```
bguess = [1 2 .5 3 1.5]
bfit = fminsearch(chisqfun,bguess)
xfit = (0:0.01:8);
yfit = ymodel(xfit,bfit);
```

Suppose that what we really care about is the area of the bump, and that the other parameters are “nuisance parameters”.

increasing temperature x in some arbitrary units
How accurately are the fitted parameters determined?
As Bayesians, we would instead say, what is their posterior distribution?

Taylor series:

\[-\frac{1}{2} \chi^2(b) \approx -\frac{1}{2} \chi^2_{\text{min}} - \frac{1}{2} (b - b_0)^T \left[ \frac{1}{2} \frac{\partial^2 \chi^2}{\partial b \partial b} \right] (b - b_0)\]

So, while exploring the $\chi^2$ surface to find its minimum, we must also calculate the Hessian (2nd derivative) matrix at the minimum.

Then

\[P(b|\{y_i\}) \propto \exp \left[ -\frac{1}{2} (b - b_0)^T \Sigma_b^{-1} (b - b_0) \right] P(b)\]

with

\[\Sigma_b = \left[ \frac{1}{2} \frac{\partial^2 \chi^2}{\partial b \partial b} \right]^{-1}\]

covariance (or “standard error”) matrix of the fitted parameters

Notice that if (i) the Taylor series converges rapidly and (ii) the prior is uniform, then the posterior distribution of the $b$'s is multivariate Normal, a very useful CLT-ish result!
Maximum Likelihood parameter errors?

Numerical calculation of the Hessian by finite difference

\[
\frac{\partial^2 f}{\partial x \partial y} \approx \frac{1}{2h} \left( \frac{f_{++} - f_{--}}{2h} - \frac{f_{+-} - f_{-+}}{2h} \right)
\]

\[
= \frac{1}{4h^2} (f_{++} + f_{--} - f_{+-} - f_{-+})
\]

\[
bfit = 1.1235 \quad 1.5210 \quad 0.6582 \quad 3.2654 \quad 1.4832
\]

\[
\text{chisqfun} = @(b) \text{sum}(((\text{ymodel}(x,b)-y)./\text{sig}).^2)
\]

\[
h = 0.1;
\]

\[
\text{unit} = @(i) (1:5) == i;
\]

\[
\text{hess} = \text{zeros}(5,5);
\]

\[
\text{for} \ i = 1:5, \ \text{for} \ j = 1:5,
\]

\[
\text{bpp} = \text{bfit} + h*(\text{unit}(i)+\text{unit}(j));
\]

\[
\text{bmm} = \text{bfit} + h*(-\text{unit}(i)-\text{unit}(j));
\]

\[
\text{bpm} = \text{bfit} + h*(\text{unit}(i)-\text{unit}(j));
\]

\[
\text{bmp} = \text{bfit} + h*(-\text{unit}(i)+\text{unit}(j));
\]

\[
\text{hess}(i,j) = (\text{chisqfun(bpp)}+\text{chisqfun(bmm)} - \text{chisqfun(bpm)}-\text{chisqfun(bmp)})./(2*h)^2;
\]

\[
\text{end}
\]

\[
\text{end}
\]

\[
\text{covar} = \text{inv}(0.5*\text{hess})
\]

This also works for the diagonal components. Can you see how?
Maximum Likelihood parameter errors?

For our example, \( y(x|b) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2} \frac{(x - b_4)^2}{b_5^2}\right) \)

\[
\begin{align*}
\text{bfit} &= \\
&= \\
&= \\
&= \\
\text{hess} &= \\
&= \\
&= \\
&= \\
\text{covar} &= \\
&= \\
&= \\
&= \\
\end{align*}
\]

This is the covariance structure of all the parameters, and indeed (at least in CLT normal approximation) gives their entire joint distribution!

The standard errors on each parameter separately are \( \sigma_i = \sqrt{C_{ii}} \)

\[
\begin{align*}
\text{sigs} &= \\
&= \\
&= \\
&= \\
\end{align*}
\]

But why is this, and what about two or more parameters at a time (e.g. \( b_3 \) and \( b_5 \))?
we have assumed that, for some value of the parameters \( b \)
the model \( y(x_i | b) \) is correct

Suppose that the model \( y(x_i | b) \) does fit. This is the null hypothesis.

Then the “statistic” \[ \chi^2 = \sum_{i=1}^{N} \left( \frac{y_i - y(x_i | b)}{\sigma_i} \right)^2 \]
is the sum of \( N \) \( t^2 \)-values.

So, if we imagine repeated experiments (which Bayesians refuse to do),
the statistic should be distributed as \( \text{Chisquare}(N) \).

If our experiment is very unlikely to be from this distribution, we
consider the model to be disproved. In other words, it is a p-value
test.
\( \chi^2 \) distribution (from Lecture 9)

\[
p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \Rightarrow \quad x \sim N(0, 1)
\]

\[y = x^2\]

\[
p_Y(y) \, dy = 2p_X(x) \, dx
\]

\[
p_Y(y) = y^{-1/2}p_X(y^{1/2}) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}
\]

\( \chi^2 \) is a “statistic” defined as the sum of the squares of \( n \) independent \( t \)-values.

\[
\chi^2 = \sum_i \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2, \quad x_i \sim N(\mu_i, \sigma_i)
\]

Chisquare(\( \nu \)) is a distribution (special case of Gamma), defined as

\[
p(\chi^2) \, d\chi^2 = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} (\chi^2)^{\frac{\nu}{2}-1} \exp\left(-\frac{1}{2} \chi^2\right) \, d\chi^2, \quad \chi^2 > 0
\]
confidence intervals

The variances of one parameter at a time imply confidence intervals as for an ordinary 1-dimensional normal distribution:

(Remember to take the square root of the variances to get the standard deviations!)

If you want to give confidence regions for more than one parameter at a time, you have to decide on a shape, since any shape containing 95% (or whatever) of the probability is a 95% confidence region!

It is conventional to use contours of probability density as the shapes (= contours of $\Delta \chi^2$) since these are maximally compact.

But which $\Delta \chi^2$ contour contains 95% of the probability?
\( \chi^2 \) distribution

Measurement precision improves with the amount of data \( N \) as \( N^{-1/2} \)

twice the data implies about twice the \( \chi^2 \) at any \( b \)

so fixed \( \Delta \chi^2 \) implies \( \sqrt{2} \) better precision
What $\Delta \chi^2$ contour in $\nu$ dimensions contains some percentile probability?

Rotate and scale the covariance to make it spherical. (Linear, so contours still contain same probability.)

Now, each dimension is an independent Normal, and contours are labeled by radius squared (sum of $\nu$ individual $t^2$ values), so $\Delta \chi^2 \sim \text{Chisquare}(\nu)$

You sometimes learn “facts” like: “delta chi-square of 1 is the 68% confidence level”. We now see that this is true only for one parameter at a time.
what is the Degree of Freedom?

How is our fit by this test?

In our example, $\chi^2(b_0) = 11.13$

This is a bit unlikely in Chisquare(20), with (left tail) $p=0.0569$.

In fact, if you had many repetitions of the experiment, you would find that their $\chi^2$ is not distributed as Chisquare(20), but rather as Chisquare(15)!

Why?

the magic word is:
“degrees of freedom” or DOF
what is the Degree of Freedom?

Degrees of Freedom: Why is $\chi^2$ with $N$ data points “not quite” the sum of $N$ $t^2$-values? Because DOF's are reduced by constraints.

First consider a hypothetical situation where the data has linear constraints:

$$t_i = \frac{y_i - \mu_i}{\sigma_i} \sim N(0, 1)$$

$$p(t) = \prod_i p(t_i) \propto \exp\left(-\frac{1}{2} \sum_i t_i^2\right)$$

$\chi^2$ is squared distance from origin $\sum t_i^2$

Linear constraint:

$$\sum_i \alpha_i y_i = C = \langle C \rangle = \sum_i \alpha_i \mu_i$$

$$C = \sum_i \alpha_i (\sigma_i t_i + \mu_i)$$

$$= \sum_i \alpha_i \sigma_i t_i + C$$

So, $\sum_i \alpha_i \sigma_i t_i = 0$

a hyper plane through the origin in t space!
what is the Degree of Freedom?

Constraint is a plane cut through the origin. Any cut through the origin of a sphere is a circle.

So the distribution of distance from origin is the same as a multivariate normal “ball” in the lower number of dimensions. Thus, each linear constraint reduces \( \nu \) by exactly 1.

We don’t have explicit constraints on the \( y_i \)’s. But as the \( y_i \)’s wiggle around (within their errors) we do have the constraint that we want to keep the MLE estimate \( b_0 \) fixed. (E.g., we have 20 wiggling \( y_i \)’s and only 5 \( b_i \)’s to keep fixed.)

So by the implicit function theorem, there are \( M \) (number of parameters) approximately linear constraints on the \( y_i \)’s. So \( \nu = N - M \), the so-called number of degrees of freedom (d.o.f.).
what is the Degree of Freedom?

Review:

1. Fit for parameters by minimizing

\[ \chi^2 = \sum_{i=1}^{N} \left( \frac{y_i - y(x_i | b)}{\sigma_i} \right)^2 \]

2. (Co)variances of parameters, or confidence regions, by the change in \( \chi^2 \) (i.e., \( \Delta \chi^2 \)) from its minimum value \( \chi^2_{\text{min}} \).

3. Goodness-of-fit (accept or reject model) by the p-value of \( \chi^2_{\text{min}} \) using the correct number of DOF.
Goodness-of-fit

Goodness-of-fit with $\nu = N - M$ degrees of freedom:

we expect $\chi^2_{\text{min}} \approx \nu \pm \sqrt{2\nu}$

this is an RV over the population of different data sets (a frequentist concept allowing a p-value)

Confidence intervals for parameters $\mathbf{b}$:

we expect $\chi^2 \approx \chi^2_{\text{min}} \pm O(1)$

this is an RV over the population of possible model parameters for a single data set, a concept shared by Bayesians and frequentists

How can $\pm O(1)$ be significant when the uncertainty is $\pm \sqrt{2\nu}$?

Answer: Once you have a particular data set, there is no uncertainty about what its $\chi^2_{\text{min}}$ is. Let’s see how this works out in scaling with $N$:

$\chi^2$ increases linearly with $\nu = N - M$

$\Delta \chi^2$ increases as $N$ (number of terms in sum), but also decreases as $(N^{-1/2})^2$, since $\mathbf{b}$ becomes more accurate with increasing $N$:

$$
\Delta \chi^2 \propto N(\delta b)^2, \quad \delta b \propto N^{-1/2} \quad \Rightarrow \quad \Delta \chi^2 \propto \text{const}
$$

quadratic, because at minimum

universal rule of thumb
linear error propagation for arbitrary function of parameters

What is the uncertainty in quantities other than the fitted coefficients:

Method 1: Linearized propagation of errors

\(b_0\) is the MLE parameters estimate

\(b_1 \equiv b - b_0\) is the RV as the parameters fluctuate

\[
    f = f(b) = f(b_0) + \nabla f \cdot b_1 + \cdots
\]

\[
    \langle f \rangle \approx \langle f(b_0) \rangle + \nabla f \cdot \langle b_1 \rangle = f(b_0)
\]

\[
    \langle f^2 \rangle - \langle f \rangle^2 \approx 2f(b_0)\langle \nabla f \cdot \langle b_1 \rangle \rangle + \langle (\nabla f \cdot b_1)^2 \rangle
\]

\[
    = \nabla f \cdot \langle b_1 b_1^T \rangle \cdot \nabla f^T
\]

\[
    = \nabla f \cdot \Sigma \cdot \nabla f^T
\]
linear error propagation for arbitrary function of parameters

In our example, if we are interested in the area of the "hump",

\[
\text{bfit} = \\
\begin{array}{ccccc}
1.1235 & 1.5210 & 0.6582 & 3.2654 & 1.4832 \\
\text{covar} = & 0.1349 & 0.2224 & 0.0068 & -0.0309 & 0.0135 \\
& 0.2224 & 0.6918 & 0.0052 & -0.1598 & 0.1585 \\
& 0.0068 & 0.0052 & 0.0049 & 0.0016 & -0.0094 \\
& -0.0309 & -0.1598 & 0.0016 & 0.0746 & -0.0444 \\
& 0.0135 & 0.1585 & -0.0094 & -0.0444 & 0.0948
\end{array}
\]

\[
f = b_3 b_5
\]

\[
\nabla f = (0, 0, b_5, 0, b_3)
\]

\[
\nabla f \Sigma \nabla f^T = b_5^2 \Sigma_{33} + 2b_3 b_5 \Sigma_{35} + b_3^2 \Sigma_{55} = 0.0336
\]

\[
\sqrt{0.0336} = 0.18
\]

so \( b_3 b_5 = 0.98 \pm 0.18 \)

A function of normals is not normal
Sampling the posterior histogram

**Method 2: Sample from the posterior distribution**

1. Generate a large number of (vector) $b$'s
   
   $$b \sim \text{MVNormal}(b_0, \Sigma_b)$$

2. Compute your $f(b)$ separately for each $b$

3. Histogram

Note again that $b$ is typically (close to) m.v. normal because of the CLT, but your (nonlinear) $f$ may not, in general, be anything even close to normal!
Sampling the posterior histogram

Our example:

```matlab
bees = mvnrnd(bfit, covar, 10000);
humps = bees(:,3).*bees(:,5);
hist(humps, 30);
std(humps)
```

```
std = 0.1833
```

Does it matter that I use the full covar, not just the 2x2 piece for parameters 3 and 5?